BAB I
PENDAHULUAN

1.1 Latar Belakang

Ketidakseimbangan data adalah masalah yang sering dijumpai pada klasifikasi. Ketidakseimbangan data terjadi saat jumlah kelas minoritas lebih kecil dibandingkan dengan jumlah kelas mayoritas [1]. Hal ini dapat menyebabkan terjadinya bias dalam mengklasifikasikan kelas minoritas dan tentunya dapat memengaruhi kinerja *machine learning* [2].

Terdapat beberapa cara untuk mengatasi ketidakseimbangan data pada klasifikasi. Penanganan ketidakseimbangan data ini dapat dikelompokkan menjadi tiga bagian, yaitu pendekatan tingkatan data, pendekatan tingkatan algoritma, dan *cost-sensitive learning* [3]. Pada tingkatan data salah satu cara untuk mengatasi ketidakseimbangan data adalah dengan menerapkan metode sampling [4].

Metode sampling adalah pendekatan untuk menyeimbangkan distribusi kelas minoritas dan kelas mayoritas. Metode ini terbagi menjadi tiga bagian, yaitu undersampling, oversampling dan kombinasi oversampling dan undersampling (hybrid sampling). Untuk menyeimbangkan data, undersampling menghapus objek pada kelas mayoritas, sehingga jumlah objek yang dimiliki setiap kelas sama. Oversampling menambahkan objek baru pada kelas minoritas. Hybrid sampling merupakan kombinasi dari oversampling dan undersampling. Metode sampling ini, menambahkan objek baru pada kelas minoritas dan menghapus objek pada kelas mayoritas [5].

Oversampling yang merupakan bagian dari pendekatan tingkat data meningkatkan jumlah sampel pada kelas minoritas untuk mengurangi rasio ketidakseimbangan data dan meminimalisir kesalahan klasifikasi [6]. Kelebihan dari oversampling adalah tidak menyebabkan hilangnya informasi pada dataset dan menurut Nabil et al, [7] oversampling cocok digunakan untuk data yang berukuran kecil.

Penelitian yang dilakukan oleh Chamidah, et al [8] menunjukkan bahwa metode oversampling efektif untuk meningkatkan performa klasifikasi pada algoritma Naïve

Bayes (NB), Decision Tree (DT), dan Artificial Neural Network (ANN) dibandingkan tanpa oversampling. Kinerja paling baik dihasilkan menggunakan algoritma ANN dengan akurasi sebesar 0,91 setelah penerapan oversampling dan 0,84 sebelum penerapan oversampling. Hasil evaluasi juga menunjukkan nilai *recall* yang tinggi setelah penerapan metode oversampling yang artinya kemampuan *classifier* untuk memprediksi kelas minoritas (kelas hipertensi) cukup tinggi. Penelitian lain menggunakan metode oversampling Adaptive Synthetic Sampling (ADASYN) dan Synthetic Minority Oversampling Technique (SMOTE) untuk menyeimbangkan data yang kemudian diklasifikasikan dengan algoritma Support Vector Machine (SVM). Penelitian tersebut membuktikan terdapat peningkatan akurasi sebesar 2-4% dibandingkan hanya menggunakan model klasifikasi Support Vector Machine (SVM) [9].

Berdasarkan beberapa penelitian sebelumnya, penelitian ini menggunakan metode Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), dan Borderline-SMOTE untuk menangani ketidakseimbangan kelas pada dataset, kemudian diklasifikasikan menggunakan algoritma random forest. Tujuan penelitian ini adalah untuk meningkatkan performa *random forest classification* pada data yang tidak seimbang dengan menerapkan metode oversampling.

1.2 Rumusan Masalah

Adapun masalah yang diangkat, dibahas dan diselesaikan adalah:

1. Bagaimana performa *random forest classification* setelah implementasi metode oversampling pada data yang tidak seimbang?
2. Metode oversampling apa yang menghasilkan performa terbaik pada *random forest classification*?

1.3 Batasan Masalah

Adapun batasan masalah pada penelitian ini, yaitu:

1. Metode sampling yang digunakan adalah ADASYN, SMOTE, dan Borderline-SMOTE.
2. Model klasifikasi yang digunakan adalah Random Forest.

3. Penelitian ini menggunakan empat dataset dari *KEEL repository*, yaitu *glass1*, *page-blocks0*, *yeast1*, dan *yeast6*.
4. Evaluasi performa menggunakan *confusion matrix*, dengan metrik yang digunakan adalah *accuracy*, *recall*, *f1-score* dan *g-mean*.

1.4 Tujuan Penelitian

1. Mengetahui performa *random forest classification* setelah implementasi metode oversampling pada data yang tidak seimbang.
2. Mengetahui metode oversampling yang paling baik untuk menangani ketidakseimbangan data pada *random forest classification*.

1.5 Manfaat Penelitian

1.5.1 Manfaat Teoritis

Penelitian ini diharapkan dapat memberikan wawasan baru mengenai implementasi metode oversampling pada data yang tidak seimbang dan pengaruhnya terhadap performa *random forest classification*.

1.5.2 Manfaat Praktis

Penelitian ini dapat menjadi referensi untuk peneliti yang ingin meneliti tentang ketidakseimbangan kelas, khususnya terkait implementasi metode oversampling dan pengaruhnya terhadap performa *random forest classification*.

1.6 Sistematika Penulisan

BAB I PENDAHULUAN, terdiri dari enam sub bab, yaitu latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA, berisi studi literatur berdasarkan penelitian sebelumnya dan uraian dasar-dasar teori yang terkait dengan penelitian.

BAB III METODE PENELITIAN, bab ini berisi penjelasan tentang tahapan penelitian yang dilakukan peneliti.

BAB IV HASIL DAN PEMBAHASAN, berisi pembahasan mengenai metode yang diterapkan dan hasil evaluasi dari klasifikasi Random Forest dengan metode oversampling.

BAB V PENUTUP, berisi kesimpulan dari hasil penelitian dan saran untuk penelitian selanjutnya.